Real-Time Multi-SLAM System for Agent Localization and 3D Mapping in Dynamic Scenarios

Pierre Alliez[1], Fabien Bonardi[2], Samia Bouchafa[2], Jean-Yves Didier[2], Hicham Hadj-Abdelkader[2], Fernando Ireta Muñoz[1], Viachaslau Kachurka[2], Bastien Rault[4], Maxime Robin[4], David Roussel[2]

***Abstract*— This paper introduces a Wearable SLAM system that performs indoor and outdoor SLAM in real time. The related project is part of the MALIN challenge which aims at creating a system to track emergency response agents in complex scenarios (such as dark environments, smoked rooms, repetitive patterns, building floor transitions and doorway crossing problems), where GPS technology is insufficient or inoperative. The proposed system fuses different SLAM technologies to compensate the lack of robustness of each, while estimating the pose individually. LiDAR and visual SLAM are fused with an inertial sensor in such a way that the system is able to maintain GPS coordinates that are sent via radio to a ground station, for real-time tracking. More specifically, LiDAR and monocular vision technologies are tested in dynamic scenarios where the main advantages of each have been evaluated and compared. Finally, 3D reconstruction up to three levels of details is performed.**

I. INTRODUCTION

Simultaneous localization and mapping (SLAM) has been one of the most studied topic in the fields of robotics and computer vision. Various applications such as autonomous navigation, indoor reconstruction and urban 3D modeling can now be adequately performed using different technologies on robotic platforms [1]–[5]. However, more complicated tasks such as search and rescue under uncontrolled conditions still require the presence of trained agents (civil security, firefighters, soldiers, etc.) to perform reckon missions in highly dynamic environments.

A Wearable SLAM System (WSS) focuses on the idea of an accurate real-time localization of the bearer in highly dynamic conditions, while also sending and registering information of the environment. To the best of our knowledge, this problem has been formulated separately as an indoor or outdoor SLAM and attempted to be solved by fusing different technologies such as inertial sensors, vision, ultrasound in case of indoor localization and GPS, radio or even terrestrial cellular networks in case of outdoor localization.

The problem of accurately localizing emergency response agents in an unknown environment remains an open problem.

[1] TITANE Project-Team, 06902 INRIA Sophia Antipolis Méditerrannée, France {fernando.ireta-munoz, pierre.alliez}@inria.fr

[2] IBISC, Univ Evry, Université Paris-Saclay, 91025, Evry, France {viachaslau.kachurka, david.roussel, hicham.hadjabdelkader, fabien.bonardi, jeanyves.didier, samia.bouchafa}@univ-evry.fr

[3] INNODURA TB, 69603, Villeurbanne, France {bastien.rault, maxime.robin}@innodura.fr

* Authors are listed alphabetically.

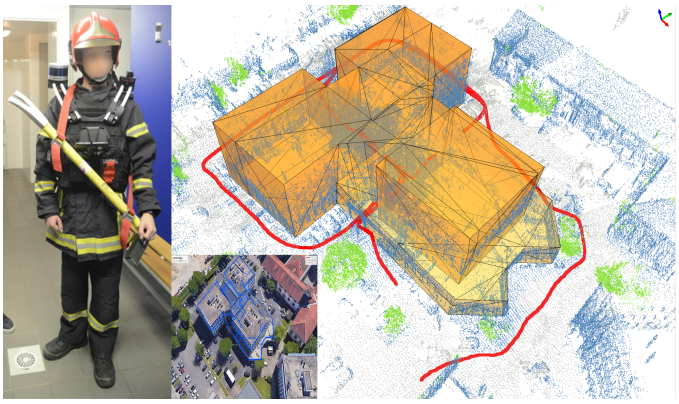

Fig. 1. Agent-based localization SLAM system. A fusion between different sensors (LiDAR, IMU, camera and GPS) is performed for achieving real-time indoor/outdoor SLAM. Left: Agent wearing the proposed system. Right: 3D Map (blue), trajectory (red) and 3D offline reconstruction obtained by the proposed system in an indoor/outdoor environment. Center: The obtained floorplan has been aligned with the 3D model of the building in Google Earth (closeup).

The main difficulty lies in the performance of localization technologies that varies according to the environment conditions and the lack of suitable technologies that can take into account the technical limits as small and efficient equipments. More reliable and robust systems have been obtained by fusing different technologies in order to pick the main advantages of each.

In this paper, a multi-sensor WSS for indoor/outdoor localization in highly dynamic environments has been developed. The proposed system (Fig. 1 *left*) has been tested under the conditions established by the MALIN challenge, which aims at accurately locating agents with or without GPS signal available. Examples of different encountered difficulties are shown in Fig. 3. Furthermore, the acquired pointclouds have been post-processed offline to produce 3D models at three different levels of detail (LOD). Several applications which require human intervention would benefit from these systems and useful in the robotics field.

The main contributions of this paper are:

- LiDAR-Visual-Inertial (LVI) fusion strategy for performing real time indoor/outdoor SLAM.
- Visual-Inertial (VI) SLAM non destructive reinitialization to recover from tracking failure.
- LVI-SLAM map to GPS UTM map registration strategy and probabilistic pose filtering.
- Offline 3D LOD reconstruction.

This paper is organized as follows: Section II is devoted to introduce the state-of-art of LVI fusion that performs real-time SLAM. The WSS and the LVI-GPS sensor fusion strategy used here are introduced in Section III. Section IV shows both, the obtained results while performing real-time localization and a post-processing offline 3D LOD reconstruction framework extended from a previous method [6]. Finally, the paper ends with a conclusion and future works.

II. LVI LOCALIZATION AND MAPPING

Real-time SLAM is currently a widely studied topic in robotics and computer vision communities. The availability of new sensors and computing power have motivated paradigm shifts in SLAM algorithms and environment representations, where large-scale fully dense maps can be obtained and post-processed for detailed 3D reconstruction.

Particularly for camera sensors, visual SLAM has provided outstanding results for accurately estimate trajectories at considerably high framerates. Various techniques based on RGB-D sensors [7]–[10], monocular [11], [12], stereo cameras [13], [14] or recent event cameras [15] and 360°cameras [9], [16], have been extensively investigated. On the other hand, LiDAR SLAM methods have introduced dense mapping with higher accuracy. The position of the sensor is estimated by using geometric-based techniques mainly based on the well known Iteratively Closest Points (ICP) method [17] and its variants [18]–[20]. Real-time LiDAR SLAM techniques can be found in [21], [22].

Visual and LiDAR SLAM share similar pipelines while estimating the pose. The main approaches are cited in [23]. In general, both strategies follow the non-linear IRLS (Iteratively Re-Weighted Least Squares) framework described below:

1) Acquire a reference and a current dataset.
2) Extract features and find their closest points for each dataset.
3) Evaluate an error between the two datasets by using robust weights.
4) Estimate pose by transforming/warping one of the datasets (current or reference).
5) Repeat to 3 until convergence.

Both, visual and LiDAR SLAM, are often paired with inertial sensors for estimating more robust and accurate 6 DOF (Degrees of Freedom) poses. The association of measurements can reduce the drift of the position while performing tracking. Visual-inertial (VI) methods can perform robust estimation of the pose when rich visual information is available. Recent VI-SLAM methods have been cited and classified based on their operating principles in [24]. Relying upon the same principles as VI-SLAM, LiDAR-Inertial (LI) SLAM has become a relevant technology for autonomous navigation and robotics [25]–[27]. The main advantages of using LiDAR are that it is not sensitive to lighting conditions, repetitive patterns or even smoke in the scene and it can register dense pointclouds at relatively high frequencies.

Real-time LVI-SLAM has been recently proposed in the literature [28]–[31] as it combines the main advantages of each technology. The fusion of the sensors has been categorized into tightly-coupled and loosely-coupled based on the dependency between the sensors for estimating the pose. Tightly-coupled methods refine the pose either by local and/or global optimization or by probabilistic filtering (EKF e.g.). In most cases IMU is used for prediction of the poses. On the other hand, loosely-coupled methods prioritize the main pose estimation process obtained by one sensor, which is aided by the poses obtained by other sensors in a separate process (e.g. Lidar or visual localization aided by IMU). The fusion of all approaches cited above use the Kalman filter (KF) and are based on the LOAM algorithm [21] for extracting geometric features with the exception of [28], which uses a dense registration on denoised pointclouds. The approaches mainly differ in the type of sensors employed, how the state vector is computed and how the features are extracted from LiDAR and camera(s).

VLOAM [29] performs sequentially coarse to fine alignments by using a VI method and a LiDAR-based matching method, respectively. The inertial sensor is used for prediction which is sent to the VI-odometry for estimating the pose. The pose is refined by matching w.r.t. the LiDAR-scan. If an image-based feature is located in the area where the laser measurements are available, depth is obtained from the laser points instead of calculating from triangulation using the previously pose estimation. The association between the image-based features and the LiDAR points is made by projecting both onto a unit sphere. By using KD-trees, the mean of the three closest laser points to a detected feature from the camera is established as its 3D coordinates by assuming a local planar patch.

LIC-Fusion [30] performs an offline LVI sensor calibration which is refined online. By employing a compressed measurement model (LiDAR + Camera residuals) for updating the KF, the edges and planar SURF features are detected from LiDAR points and tightly fused along with the visual FAST features, which are extracted from the image using Kanade Lucas Tomasi (KLT) optical flow. W.r.t. other approaches cited here, this method lacks the loop closure detection stage for maintaining a global map.

VIL-SLAM [31] employs a stereo camera system for aiding frame-to-frame KLT tracking of stereo ORB features. If the number of stereo matches is below a threshold, a Shi-Tomasi Corner detector is used and ORB descriptors are computed on these features. Visual ORB features and IMU measurements are tightly fused for estimating the pose which is used for transforming current LiDAR points. Edge and Planar points are extracted from each set of current LiDAR measurements and minimized w.r.t. the generated map using the (Levenberg-Marquardt) LM-ICP algorithm for pose refinement.

Finally, [28] fuses a stereo camera with LiDAR and an IMU unit that includes two horizontal accelerometers and one vertical gyro for localization only. The method improves upon odometry by compensating the accelero. and gyro. biases that degrade velocity and position. Visual odometry provides an initial transformation for LiDAR odometry which

refines the pose using the Generalized ICP algorithm [18] over LiDAR inlier points. The azimuth error caused by uncompensated vertical gyro. bias is the main source of error. Therefore, the obtained forward velocity and azimuth are integrated with reduced IMU in KF to provide a navigation solution for urban environments. Visual features are detected using Harris corner detection and tracked using KLT. A local optimization block improves the image-based pose by eliminating outliers, and bundle adjustment is used when matched features are present across more than two frames.

Maintaining a globally consistent map throughout time and its detailed reconstruction is part of the objectives in the MALIN challenge. Loop closure detection, pose-graph optimization and deformation graph optimization are widely used for correcting both, the estimated poses and 3D map, in real time. Furthermore, global ICP approaches can guarantee a global consistent map. However, these approaches are compute-intensive due to an exhaustive search of the optimal solution in the transformation space [32]. After estimating a well-aligned pointcloud, different strategies of surface reconstruction [33] can be employed. For the purpose of this paper, the global optimization methods and surface reconstruction stages used during the competition will be briefly described here. More details are provided in the associated video.

III. WEARABLE SLAM SYSTEM

In order to introduce the proposed Wearable SLAM system, LiDAR-based and camera-based technologies for estimating the 6DOF pose will first be explained, then the fusion of these two technologies along with the inertial sensor will be shown. As mentioned in Section II, both technologies share similar pose estimation frameworks. Both approaches attempt to solve the following non-linear error function between matching measurements $\mathbf{M}^*$ and $\mathbf{M}$, where the measurements can represent here either, 3D points or intensities:

$$\mathbf{T}(\mathbf{x}) = \underset{\mathbf{T}(\mathbf{x})}{argmin} \sum_{i=1}^{N} ||\rho_i \left(\mathbf{M}_i^* - f\left(\mathbf{T}(\mathbf{x}), \mathbf{M}_i\right)\right)||^2 \in SE(3) \quad (1)$$

N denotes the number of matches and $f\left(\mathbf{T}(\mathbf{x}), \mathbf{M}_i\right)$ denotes the function that transforms a set of measurements $\mathbf{M}_i$ with transformation $\mathbf{T}(\mathbf{x})$. ρ_i is the weighting value obtained by robust estimation. The 6DOF pose $\mathbf{x} \in \mathbf{se}(3)$ can be decomposed into rotational $\mathbf{R}(\mathbf{x}) \in \mathbb{SO}(3)$ and translational $\mathbf{t}(\mathbf{x}) \in \mathbb{R}^3$ components. The group $\mathbb{SE}(3)$ has an associated Lie algebra $\mathbf{se}(3)$, comprising two separate 3-vectors $\boldsymbol{\omega}$ and $\boldsymbol{v}$ which determines the rotation and translation, respectively. The homogeneous transformation matrix $\mathbf{T}(\mathbf{x})$ has the closed-form using the exponential map as $\mathbf{T}(\mathbf{x}) = e^{[\mathbf{x}]_\wedge} \in \mathbb{R}^{4\times4}$, where $[\cdot]_\wedge$ is the twist matrix operator that can be written as $[\mathbf{x}]_\wedge = \begin{bmatrix} [\boldsymbol{\omega}]_\times & \boldsymbol{v} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{4\times4}$.

The main concept of both approaches can be summarized in two main steps:

- Find and match features between datasets.
- Compute the transformation that minimizes the distance between corresponding points.

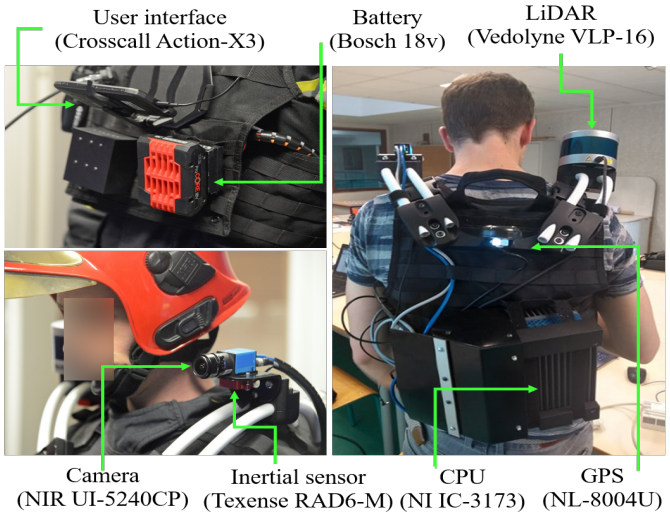

Fig. 2. Tactical waistcoat hardware configuration.

For the experiments of this paper, the 3D-3D correspondences finding is performed by using the nanoflann library [34] in case of LiDAR SLAM. The CERES library is used to solve the non-linear LM-ICP error function (2) and compute a robust estimation of the pose. In case of visual-SLAM, ORB features are triangulated from multiple views which produces a map containing keyframes and sparse points optimized using g^2o library [35]. This visual map is used for navigation but is not suitable for reconstruction due to the low points density. Both strategies are briefly described in Section III-B.3 and III-C.

A. System overview

The MALIN challenge aims to create an autonomous localization and mapping system with high portability, and without compromising the agent's mobility. Therefore, the selection and configuration of the hardware is a crucial factor.

The study of this paper has been validated on, but not limited to the hardware described in Fig. 2. The hardware setting is installed on a tactical waistcoat, where the LiDAR, camera and inertial sensors are placed on the shoulders. The autonomy of the system is about 45 minutes while performing SLAM. As a technical contribution, VI-SLAM and LI-SLAM approaches described in this paper have been compiled under Windows and used by LabView, which handles the communication between the system and the ground station via radio. Estimated latitude and longitude by the WSS are sent at 0.5 Hz.

In order to estimate the extrinsics LVI, an offline calibration strategy based on a scale ICP registration has been proposed here. First, LI and VI pose estimation processes were performed simultaneously on the same scenario and closest temporal poses are matched using their timestamps. Then, the geometric error between the points of the estimated trajectories is obtained by the Horn's method [36]. The estimated transformation that best aligns the trajectories is used as the relative pose between the sensors.

a) b) c) d) e) f)

Fig. 3. Main difficulties presented during the MALIN challenge and tested by the proposed WSS system. a) Crawling, b) Presence of rigid objects, c) Dark environment and non-rigid objects, d) Bending down and outdoor environment, e) Smoked room and f) High luminosity.

B. LiDAR-Inertial SLAM

The Velodyne VLP-16 sensor register poinclouds composed by a set of 16 beam lines at different elevation angles. Each j-th beam line contains a set of consecutive n-th 3D Euclidean points referred to as: $\{\mathbf{P}_i^j \in \mathbb{R}^3 | i < n \ \& \ j < 16\}$. The LiDAR-based method employed here is a variant of [21] with an IMU that has been integrated for improving the estimation of the pose.

The algorithm consists of three main functions:

1) Feature extraction: Similar to [29]–[31], valid feature points are extracted for each beam line and categorized into either Edge or Planar points depending on the evaluation of the geometric constraints between 2 line segments, which are constructed by linear least squares fitting (LLSF) of the k-nearest neighbors to the left and the right of each point $\mathbf{P}_i^j$ via PCA. The estimated curvature values for each geometric constraint (sharpness, large depth gap and saliency of the points) are sorted and those points with a value greater or lower than an established threshold are considered as Edges or Planar features, respectively. Compared to the original LOAM algorithm, Blob features (neither Planar nor Edge features) are not extracted since they tend to increase processing time as well as map size.

2) Local pose estimation: Depending on the labeling of features (Planar or Edge point), the pose $\mathbf{T}(\mathbf{x})$ is calculated by minimizing the Point-to-line ICP or the Point-to-point ICP error function, respectively. Both minimizations can be summarized in the following equation:

$$\mathbf{T}(\mathbf{x}) = \underset{\mathbf{T}(\mathbf{x})}{argmin} \sum_{i=1}^{N} \left|\left| \left(\mathbf{M}_i^{\omega} - \mathbf{M}^*\right)^{\top} \mathbf{A} \left(\mathbf{M}_i^{\omega} - \mathbf{M}^*\right) \right|\right|^2 \in SE(3) \tag{2}$$

where $\mathbf{M}_i^{\omega} = \mathbf{R}(\mathbf{x})\mathbf{M_i} + \mathbf{t}(\mathbf{x})$ represents here the 3D warped point. Semi-distance matrices are $\mathbf{A} = \left(\mathbb{I} - \mathbf{E}\mathbf{E}^{\top}\right)$ and $\mathbf{A} = \mathbf{E}\mathbf{E}^{\top}$ for Edges and Planar points, respectively. $\mathbf{E}$ is the eigenvector of the covariance matrix of $\mathbf{M}^* \in \mathbb{R}^3$ and $\mathbb{I} \in \mathbb{R}^{3\times3}$ is the identity matrix.

3) Global mapping optimization: A global optimization for increasing the accuracy of the estimated poses is performed. In a different thread, a Point-to-model ICP obtains the absolute pose of the current pointcloud. Each current features are matched w.r.t. voxels V_i which contain a subset of the generated 3D map. This map is then updated with the new features and filtered.

The filter is the key to maintain a real-time SLAM. A multi-barycentric method adapts the density (number of feature points per meter) for both, Edges and Planar points, according to their bounding box dimensions. The filter sets a lower or a higher density for each axis (X, Y, Z) depending on wide or narrow enviroments, respectively. This allows to perform an accurate real-time localization. The localization is obtained within 50 to 100 ms / cycle compared to the 200 ms / cycle of the original LOAM algorithm.

C. Visual-Inertial SLAM

VI-based localization has many variants [24]. For the purpose of this paper, the monocular VI-ORB-SLAM approach proposed in [37] has been employed. One of the main advantages is that its inertial component ensures scale determination with a single camera and it includes both, a loop closing and a relocalization method [38] built on ORB points descriptors of each keyframe. In addition, points belonging to moving objects in the scene are also eliminated. The tracking operation consists of two steps:

1) Match the feature points of the current frame with the map points tracked during the last frame to determine the relative pose.
2) Update and optimize local keyframes, map points and the current pose.

The main difference between ORB-SLAM [39] and VI-ORB-SLAM [37] tracking lies in the pre-integration of IMU data to provide a first estimate of the current frame's pose with respect to the last frame (step 1) or with respect to the last keyframe (step 2). Bootstrapping the VI tracking also requires a purely visual tracking during the estimation of both, the gravity vector, accelero. and gyro. biases and the scale factor based on IMU data. Since it can suddenly displace the current pose, those variables are also re-estimated during 1s each time a loop closure relocation occurs. Uncontrolled environment conditions (e.g. doorway crossings, high luminosity changes) led to a visual tracking failure. Therefore, the algorithm has been customized here with re-initialization capabilities in order to restart tracking in a non destructive way where: 1) Re-initialization preserves any already acquired map (map points + keyframes) and 2) IMU pre-integration is used to predict the current pose during visual tracking re-initialization, leading to a more accurate re-initialized pose than the motion model proposed in [40].

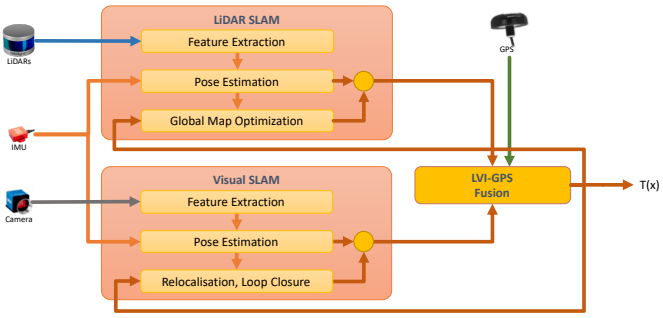

Fig. 4. Fused Agent-based pose estimation. The LiDAR-camera fusion is managed via the Kalman Filter.

D. LVI-Fusion

The proposed wearable SLAM system is concerned by the loosely-coupled fusion between LVI-SLAM method and GPS localization. Particularly, when GPS is not available, LI-SLAM receives loop closing events in VI-SLAM and VI-SLAM receives the global optimized pose from LI-SLAM. Both SLAM approaches obtain the pose at different frequencies (10Hz and 15-20Hz, respectively). Therefore, the last estimated poses from each SLAM strategy is registered at 5 Hz and sent via radio to the ground-station every 2 s.

While performing real-time reckoning missions, GPS data are maintained until its precision is over a given threshold. Only valid GPS coordinates are associated with the nearest SLAM positions using timestamps. The absolute position (in the UTM coordinates frame) between them is obtained by KF. The estimated orientation is used in the Kalman filter for prediction of the poses. Predicted GPS coordinates are used to correct the potential drift generated by LVI-SLAM. Furthermore, corrected GPS positions are sent back to both, LI-SLAM and VI-SLAM for improving the re-localisation process. Fig. 4 and 5 shows the complete fusion scheme.

E. Cartography

The offline 3D reconstruction framework used here aims to perform urban reconstruction at different LODs as in [6], but customized to indoor maps. In LOD0, all points belonging to walls, ceilings and floors are detected via semantic classification [41]. Walls are then projected onto the fitted planes of the floors. Outliers are eliminated from the projected points in order to generate a floorplan of the buildings. For LOD1, the watertight surface that delineates the building is reconstructed via a kinetic approach that computes and filters a sparse 3D arrangement of planes obtained by a kinetic computational geometry approach [42]. Finally, for LOD2, indoor 3D reconstruction is performed and concatenated with its associated LOD0 and LOD1. Results for the acquired maps here are shown in Fig. 6(d).

IV. RESULTS

In order to present the results of the proposed wearable SLAM system, part of acquired data from a set of 8 reckoning mission (ranging from 10 to 30 minutes of duration) in an outdoor/indoor environment has been used (shown in

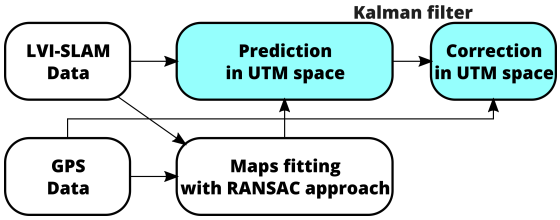

Fig. 5. Scheme of Kalman filter (in blue) used for LVI-GPS fusion.

Fig. 6(a)). The followed path offers multiple indoor/outdoor transitions, two opportunities for loop closure (one inside and one outside of a building) as well as an opportunity to assess the drift by reusing the first entry to exit the building after a long trajectory (in this case $185m$). LI-Fusion have reached the first entry with a $75cm$ drift (0.41%), whereas VI-Fusion features a $1.62m$ (0.88%) drift. Fig. 6(a)(b) shows the trajectory and map points obtained by the methods presented in Sections III-B and III-C, respectively.

In order to provide a groundtruth for both, visual and numerical comparisons, 28 GPS landmarks corresponding to building's corners have been extracted from French land register[1], which provides accurate WGS84 GPS coordinates of projected floorplans. The metric 3D coordinates of the estimated LOD0 reconstruction are converted to its WGS84 GPS coordinates and aligned w.r.t. the land register's model (See Figure 6(e)). For the groundtruth floorplan, the distance between consecutive landmarks is calculated and compared w.r.t. the length of detected lignes (by using the mehod in [43]) in case of LOD0 (estimated floorplan), and by measuring the length of the detected planes in case of LOD1.

V. CONCLUSION AND FUTURE WORK

A WSS that perform LiDAR-Visual-Inertial-GPS fusion for real-time indoor/outdoor localization and an offline 3D reconstruction has been presented. The WSS has been tested under the conditions of the MALIN challenge. Even if the LVI-fusion performs as expected with a relative small drift, various issues were encountered. An illuminator is still necessary to deal with total darkness. Current experiments are being carried out with more performant and robust visual tracking approaches such as VINS-Fusion [44]. A better performance in the smoke (up to 5 m) when using a NIR camera only has been noticed (even if a SWIR camera, which is considerably more expensive, would be more suitable). Alternate loop closure detection methods based upon LiDAR data are being investigated. The 3D cartography step requires computing oriented normals. Real-time normal estimation from LiDAR data is being explored with the aim to add oriented normals to the generated pointcloud and perform an automatic 3D LOD reconstruction after each reckoning mission. In the MALIN challenge, the 3D reconstruction stage is performed offline within 10 minutes. In the same manner, normals can be used for both, detecting more detailed edges and planar features [45] in LiDAR pointclouds and for estimating the pose using Point-to-plane ICP.

[1] https://cadastre.gouv.fr

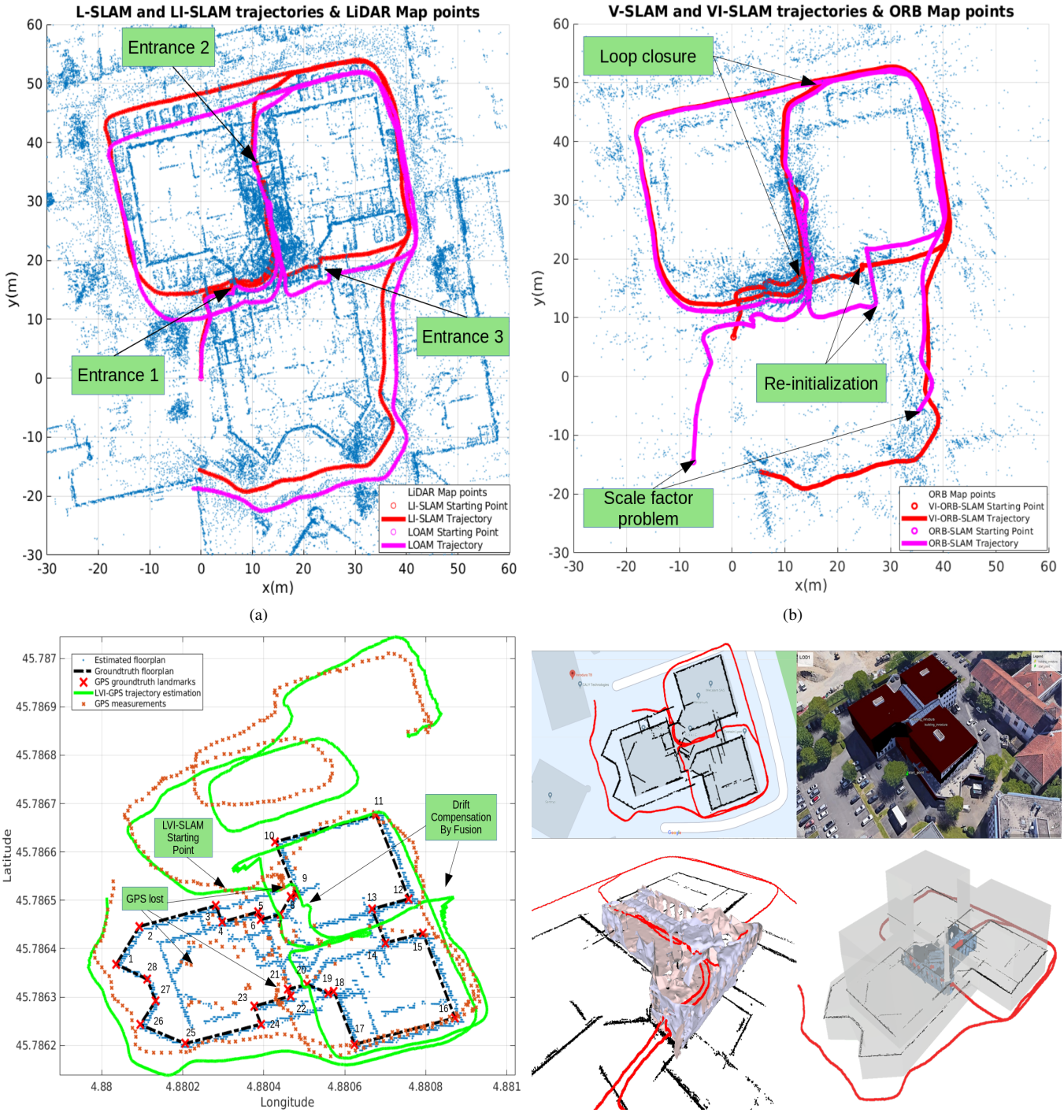

Fig. 6. Results obtained by the Wearable SLAM system. The sequence (path shown in red) consists of the following steps: 1. Walking towards Entry 1 (E1). 2. Entering building through E1 and walking along an internal corridor towards E2. 3. leaving the building from E2 and making a quarter turn around the building for reaching E3. 4. Entering from E3 and walking towards E1. 5. Leaving building from E1 and making a turn around the building to reach the end point. Re-entering the building through the entrances, had led to individual tracking loss for individual SLAM strategies. However, the tracking has been re-initialized using IMU data. a) LOAM [21] and LI-SLAM trajectories, with the globally consistent 3D map. b) ORB-SLAM [39] and VI-SLAM [37] trajectories with the sparse 3D map. c) Record of online GPS positions estimated in the UTM coordinates, where it can be clearly noticed the degraded performance for indoor tracking using only GPS. The prediction of the LVI-GPS fusion can maintain reliable GPS coordinates. Before the starting point of the LVI-SLAM, an outdoor pre-sequence is obtained for finding the relative position between the system and the GPS by using the calibration process mentioned in III-A. d) 3D LOD reconstruction results. *Top left*: LOD0 and its scale comparison w.r.t. its Google maps coordinates. *Top right*: LOD1 and its aligment w.r.t. the 3D model of the building provided in Google earth. *Bottom left:* LOD3 of the indoor visited scenario, obtained by Poisson reconstruction. *Bottom right:* Outdoor and indoor LOD reconstruction are shown together.

ACKNOWLEDGMENTS

This work takes part as the LOCA3D project in the framework of the challenge MALIN funded with the support of Directorate General of Armaments and French National Research Agency *https://challenge-malin.fr*.

REFERENCES

[1] J. A. Castellanos, J. Martinez, J. Neira, and J. D. Tardos, "Simultaneous map building and localization for mobile robots: A multisensor fusion approach," in *Proceedings of the 1998 IEEE International Conference on Robotics and Automation*, vol. 2. IEEE, May 1998, pp. 1244–1249.

[2] D. Nistér, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1. IEEE, July 2004, pp. I–I.

[3] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, June 2014, pp. 15–22.

[4] G. Nützi, S. Weiss, D. Scaramuzza, and R. Siegwart, "Fusion of IMU and vision for absolute scale estimation in monocular SLAM," *Journal of Intelligent & Robotic Systems*, vol. 61, no. 1, pp. 287–299, Jan. 2011.

[5] J. Levinson and S. Thrun, "Robust vehicle localization in urban environments using probabilistic maps," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2010, pp. 4372–4378.

[6] Y. Verdie, F. Lafarge, and P. Alliez, "Lod generation for urban scenes," *ACM Trans. Graph.*, vol. 34, no. 3, May 2015.

[7] F. I. Ireta Muñoz and A. I. Comport, "Point-to-hyperplane icp: Fusing different metric measurements for pose estimation," *Advanced Robotics Journal*, September 2017.

[8] C. Kerl, J. Sturm, and D. Cremers, "Dense visual slam for rgb-d cameras," in *IROS*, 2013.

[9] M. Meilland and A. Comport, "On unifying key-frame and voxel-based dense visual slam at large scales," in *International Conference on Intelligent Robots and Systems*. Tokyo, Japan: IEEE/RSJ, 3-8 November 2013.

[10] T. Whelan, S. Leutenegger, R. S. Moreno, B. Glocker, and A. Davison, "Elasticfusion: Dense slam without a pose graph," in *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.

[11] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *ECCV*, September 2014.

[12] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 1052–1067, 2007.

[13] J. Engel, J. Stueckler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *IROS*, Sept. 2015.

[14] T. Pire, T. Fischer, G. Castro, P. De Cristóforis, J. Civera, and J. Berlles, "S-ptam: Stereo parallel tracking and mapping," *Robotics and Autonomous Systems*, vol. 93, Apr. 2017.

[15] H. Rebecq, T. Horstschaefer, G. Gallego, and D. Scaramuzza, "Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real-time," *IEEE Robotics and Automation Letters*, vol. PP, Dec. 2016.

[16] M. Meilland, A. I. Comport, and P. Rives, "A spherical robot-centered representation for urban navigation," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2010, pp. 5196–5201.

[17] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 14, no. 2, pp. 239–256, Feb 1992.

[18] A. Segal, D. Hähnel, and S. Thrun, "Generalized-icp," June 2009.

[19] Y. Chen and G. Medioni, "Object modeling by registration of multiple range images," in *IEEE International Conference on Robotics and Automation*, Sacramento, CA, USA, Apr 1991.

[20] J. Serafin and G. Grisetti, "Nicp: Dense normal based point cloud registration," in *IROS*, Hamburg, Germany, 2015, pp. 742–749.

[21] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time," in *Proceedings of the Robotics: Science and Systems*, July 2014.

[22] J. Behley and C. Stachniss, "Efficient surfel-based slam using 3d laser range data in urban environments," June 2018.

[23] H. Huang, J. Zhao, and J. Liu, "A survey of simultaneous localization and mapping," Aug. 2019.

[24] C. Chang, H. Zhu, M. Li, and S. You, "A review of visual-inertial simultaneous localization and mapping from filtering-based and optimization-based perspectives," *Robotics*, vol. 7, p. 45, Aug. 2018.

[25] H. Ye, Y. Chen, and M. Liu, "Tightly coupled 3d lidar inertial odometry and mapping," *2019 International Conference on Robotics and Automation (ICRA)*, May 2019.

[26] C. Qin, H. Ye, C. E. Pranata, J. Han, S. Zhang, and M. Liu, "R-lins: A robocentric lidar-inertial state estimator for robust and efficient navigation," 2019.

[27] C. Park, P. Moghadam, S. Kim, A. Elfes, C. Fookes, and S. Sridharan, "Elastic lidar fusion: Dense map-centric continuous-time slam," 2017.

[28] Y. Balazadegan, S. Hosseinyalamdary, and Y. Gao, "Visual-lidar odometry aided by reduced imu," *ISPRS International Journal of Geo-Information*, vol. 5, p. 3, Jan. 2016.

[29] J. Zhang and S. Singh, "Laser-visual-inertial odometry and mapping with high robustness and low drift," *Journal of Field Robotics*, Aug. 2018.

[30] X. Zuo, P. Geneva, W. Lee, Y. Liu, and G. Huang, "Lic-fusion: Lidar-inertial-camera odometry," 2019.

[31] W. Shao, S. Vijayarangan, C. Li, and G. Kantor, "Stereo visual inertial lidar simultaneous localization and mapping," 2019.

[32] F. I. Ireta Muñoz and A. I. Comport, "Global point-to-hyperplane icp: Local and global pose estimation by fusing color and depth," in *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Daegu, South Korea, 2017.

[33] M. Berger, A. Tagliasacchi, L. Seversky, P. Alliez, G. Guennebaud, J. Levine, A. Sharf, and C. Silva, "A survey of surface reconstruction from point clouds," *Computer Graphics Forum*, pp. n/a–n/a, Mar. 2016.

[34] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *International Conference on Computer Vision Theory and Application VISSAPP'09)*. INSTICC Press, 2009, pp. 331–340.

[35] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2o: A general framework for graph optimization," in *2011 IEEE International Conference on Robotics and Automation (ICRA 2011)*, May 2011, pp. 3607–3613.

[36] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[37] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, p. 796–803, Apr 2017.

[38] D. Gálvez-López and J. D. Tardós, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[39] R. Mur-Artal, J. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[40] V. Kachurka, D. Roussel, H. Hadj-Abdelkader, F. Bonardi, J.-Y. Didier, and S. Bouchafa, "Swir camera-based localization and mapping in challenging environments," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 446–456.

[41] F. Lafarge and C. Mallet, "Creating large-scale city models from 3d-point clouds: A robust approach with hybrid representation," *International Journal of Computer Vision*, vol. 99, Aug. 2012.

[42] J.-P. Bauchet and F. Lafarge, "Kinetic shape reconstruction," *ACM Trans. Graph.*, vol. 39, no. 5, June 2020. [Online]. Available: https://doi.org/10.1145/3376918

[43] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," *Comput. Graph. Forum*, vol. 26, pp. 214–226, June 2007.

[44] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, Aug. 2018.

[45] W. S. Grant, R. C. Voorhies, and L. Itti, "Finding planes in lidar point clouds for real-time registration," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov 2013, pp. 4347–4354.